**Data and Method**

**Methodology**

The research, as already explained, is mainly composed of two different parts: the first, that uses unsupervised learning to obtain clusters of citizens’ attitudes towards climate change, and the second, that uses supervised learning to predict pro-environment behaviour.

The first set of methods focuses on identifying some profiles of citizens’ attitudes towards climate change using different types of unsupervised learning techniques: K-means clustering and Correlational Class Analysis (CCA).

K-means clustering is a “numerical, unsupervised, non-deterministic, iterative method” (Na et al., 2010, p.63) and it seeks to identify a finite set of clusters or subgroups to describe data (Fonseca, 2013; James et al., 2013). This method creates some subgroups in order to maximize both the similarity within clusters and the differences among other groups

Correlational Class Analysis (CCA) identify such “cultural schemas” in a survey data, in particular in a public opinion data (Boutyline, 2017; Rossoni et al., 2020). This technique is an implementation of Relational Class Analysis (RCA) developed by Goldberg (2011) and it “seeks to parse out groups, or classes, of like-minded individuals. Unlike these methods, however, it uses relationality to compare these individuals not on their attitudes per se but on the patterns of relations between their attitudes” (p.1399). Therefore, the goal of RCA is to partition individuals into groups which shared “cultural classes” (Rossoni et al., 2020).The shared “cultural schemas” “does not imply having identical attitudes or behaviours, rather it suggests being in agreement on the structures of relevance and opposition that make actions and symbols meaningful” (Goldberg, 2011, p.1402). Therefore, it tries to find patters of associations between attitudes or behaviours in terms of “relationality”. In addition, it tries to find relationships both between individuals and between variables, combining clustering analysis and multidimensional scaling or factor analysis (Goldberg, 2011). The difference between RCA and CCA lies in the concept of “relationality”. In fact, while Goldberg (2011) uses linear dependency between two individuals vectors of responses in order to find the shared cultural schemas, CCA suggests to adopt Pearson’s correlation (Boutyline, 2017). Boutyline (2017) demonstrated that CCA produces more accurate results.

Using these algorithms only quantitative variables can be used, in fact only climate change questions are considered, except for the dependent variable, pro-environmental action, and climate change risk perception.[[1]](#footnote-1) The purpose of this part of analysis is to find some different types of citizens, called clusters or classes, that better describe the data used. In fact, through these techniques some new segmentations of citizens could be identified and then they could help to find new explanations to the phenomenon studied. Theoretically, using these two different types of segmentations of citizens, the results should be opposite. On the one side, the traditional clustering profiles the data according to similar attitudes. On the other side, CCA finds shared cultural schemas, structure of thought. Eventually, the classes obtained from k-means clustering and correlational class analysis are used as predictors in the subsequent classifications.

The second set of methods focuses on prediction climate change pro-environment using different types of supervised learning techniques and classifiers. In fact, classification is used when a categorical variable is predicted (James et al., 2013). “The methods used for classification first predict the probability of each of the categories of a qualitative variable” (James et al., 2013, p. 127).

The different techniques of classifiers are briefly presented as follow.

The action prediction starts with a Logistic Regression. It is a form of binary regression and it explains relationships between a categorical outcome and some continuous or discrete predictors (Peng et al., 2002). It models the probability of being to a particular category (Peng et al., 2002; Stoltzfus, 2011).

The model requires some assumptions:

1. independence of errors;
2. linearity in the logit for continuous independent variables;
3. the absence of multicollinearity among explanatory variables;
4. the absence of extreme outliers

(Stoltzfus, 2011)

However, some assumptions are violated. In fact, there is no present linearity in the logit for age variable. In addition, some outliers are found in climate change risk perception, but they are not so far away from the rest of the value.

In spite of the robustness of the logistic regression models, data cannot fully satisfy the assumptions, also decision tree models are fitted. Decision Tree is a “flow-chart-like hierarchical tree structure” (Jenhani et al., 2008, p. 786) and it is composed of three elements: nodes, edges and leaves. Nodes represent attributes or variables, edges correspond to the different possible attribute values and lastly leaves include objects that typically belong to the same class or that are very similar (Jenhani et al., 2008). The main advantages of decision tree are that it has not assumptions and especially it produces graphical representation, which make it easier to read and to interpret the model.

The analysis continues with another robust model: Random Forest, which is produces of multiple and randomized decision trees that operate as an ensemble (Belgiu, 2016; Biau & Scornet, 2016). This classifier “can successfully handle high data dimensionality and multicollinearity, being both fast and insensitive to overfitting” (Belgiu, 2016, p.24). Another advantage is that it can dealing with unbalanced data, as in this case (Belgiu, 2016).

The last classifier used is Gradient Boosting. It is similar to random forest algorithm, but this case each new tree is been created using the previous ones, in order to correct mistakes made (James et al., 2013). Instead of fitting a large amount of trees separately, it learns slowly by previous trees recursively.

The last two tree-based methods, producing multiple trees, have become more popular since they improve in prediction accuracy but they loss in interpretation (Belgiu, 2016; James et al., 2013).

To sum up, all these classifiers have the possibility to predict pro-environmental behaviour. Socio-demographic variables, classes created form k-means and CCA and climate change risk perception are used as predictors. In fact, we want also to investigate the main factors and predictors that influence pro-environmental behavior. This process is achieved thanks to selected models, logistic regression and tree-based methods, which can determine the importance of independent variables.

**Data Description**

As aforementioned, the research studies pro-environmental behaviour of European citizens. The main data used in this project is the Special Eurobarometer 91.3 dataset made available by the Eurobarometer Open Data website. The Eurobarometer is a public opinion research institution in the European Union with the aim to examine a variety of topics and attitudes. In this survey, entitled “Climate Change”, there are 27655 respondents from 28 countries of the European Union. The Eurobarometer data are available from GESIS (European Commission, Brussels, 2019). Eurobarometer 91.2 asks some questions about environmental issues and some socio-demographic information. Some relevant questions about climate change and socio-demographic variables are selected, as shown in the table 1.

TAB. 1: Variables selected

|  |  |  |
| --- | --- | --- |
| **Questions** | **Type** | **Code** |
| ***Question about Climate Change issues*** |  |  |
| Have you personally taken any action to fight climate change over the past six months? | Dummy | qb5 |
| And how serious a problem do you think climate change is at this moment? Please use a scale from 1 to 10, with '1' meaning it is "not at all a serious problem" and '10' meaning it is "an extremely serious problem" | 10 points Likert scale | qb2 |
| To what extent do you agree or disagree with each of the following statements? Taking action on climate change will lead to innovation that will make EU companies more competitive | 4 points Likert scale | qb4\_3 |
| To what extent do you agree or disagree with each of the following statements? Adapting to the adverse impacts of climate change can have positive outcomes for citizens in the EU (N) | 4 points Likert scale | qb4\_5 |
| How important do you think it is that the (NATIONALITY) government sets ambitious targets to increase the amount of renewable energy used, such as wind or solar power, by 2030? | 4 points Likert scale | qb7 |
| How important do you think it is that the (NATIONALITY) government provides support for improving energy efficiency by 2030 (e.g. by encouraging people to insulate their home or buy electric cars)? | 4 points Likert scale | qb8 |
| To what extent do you agree or disagree with the following statement: We should reduce greenhouse gas emissions to a minimum while offsetting the remaining emissions, for instance by increasing forested areas, to make the EU economy climate neutral by 2050. | 4 points Likert scale | qb9 |
| ***Socio-demographic information*** |  |  |
| In political matters people talk of "the left" and "the right". How would you place your views on this scale? | 10 points Likert scale | d1 |
| Which of the following best corresponds to your own current situation? | Categorical | d7 |
| How old were you when you stopped full-time education? | Continuous | d8 |
| Gender | Dummy | d10 |
| How old are you? | Continuous | d11 |
| Would you say you live in a...? | Categorical | d25 |
| Do you see yourself and your household belonging to…? | Categorical | d63 |
| Country | Categorical | country |

**Data Cleaning**

The first step before performing the analysis is data cleaning. In order to obtain an accurate analysis some observations are dropped. In fact, missing data or refusal answers of climate change issues are not considered in the final dataset. The reason why k-means clustering and CCA does not accept missing data and therefore the entire observation must be removed. Instead, socio-demographic variables are for the most part categorical and then *refusal* or *dk* (don’t know) are kept among the answer choices. However, some transformations are adopted in these variables. Political orientation (encoded as d1) is originally presented in a 10-points Likert scale (1 left and 10 right). It is transformed in a categorical variable: the answers 1-2 are become “left”, 3-4 “centre-left”, 5-6 “centre”, 7-8 “centre-right”, 9-10 “right” and dk or refusal “not positionable”. For the current situation variable (d7), some new categories are created, depending on whether individual has declared that he/she lives with “partner”, “partner and children” or he/she is “single” or “single (and he/she lives) with children”. The education variable (d8), or rather when he/she finished education, is been converted from continuous to categorical. According to scholars (Abu-Omar & Rütten, 2008; Loyen, 2016) five categories are created: “up to 15 years”, “16-19 years”, “20+years”, “still studying” and “refusal/other”. Gender (d10) and age (d11) are not manipulated since nobody answered with “other” and therefore the first variable is a dummy encoded as “male” and “female”, while the second one is a continuous. For place of residence (d25) and class identity (d63) variables, the categories proposed by the Eurobarometer are kept. The first variable has the following classes: “rural area or village”, “small or middle sized town”, “large town” and “dk” (don’t know). While class identity variable has: “the working class of society”, “the lower middle class of society”, “the middle class of society”, “the upper middle class of society”, “the higher class of society”. Table 2 presents summary statistics for each of these variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Code** | **Obs** | **Mean** | **Std. dev.** | **Min** | **Max** |
| Qb5 |  |  |  |  |  |
| *Yes* | *14327* |  |  |  |  |
| *No* | *7651* |  |  |  |  |
| qb2 | 21978 | 7.93 | 2.02 | 1 | 10 |
| qb4\_3 | 21978 | 1.74 | 0.71 | 1 | 4 |
| qb4\_5 | 21978 | 1.90 | 0.87 | 1 | 4 |
| qb7 | 21978 | 1.52 | 0.65 | 1 | 4 |
| qb8 | 21978 | 1.56 | 0.68 | 1 | 4 |
| qb9 | 21978 | 1.50 | 0.62 | 1 | 4 |
| d1 | 21978 |  |  |  |  |
| *left* | *1853* |  |  |  |  |
| *Centre-letf* | *3856* |  |  |  |  |
| *centre* | *7968* |  |  |  |  |
| *Centre-right* | *3470* |  |  |  |  |
| *Right* | *1603* |  |  |  |  |
| *Not positionable* | *3228* |  |  |  |  |
| d7 | 21978 |  |  |  |  |
| *partner* | *7791* |  |  |  |  |
| *Patner and children* | *7000* |  |  |  |  |
| *Single* | *5975* |  |  |  |  |
| *Single with children* | *1120* |  |  |  |  |
| *Refusal/Other* | *92* |  |  |  |  |
| d8 | 21978 |  |  |  |  |
| *up to 15 years old* | *2665* |  |  |  |  |
| *16-19 years old* | *10013* |  |  |  |  |
| *20+ years old* | *8981* |  |  |  |  |
| *Refusal/dk* | *319* |  |  |  |  |
| d10 | 21978 |  |  |  |  |
| *Man* | *10527* |  |  |  |  |
| *Woman* | *11451* |  |  |  |  |
| d11 | 21978 | 50.51 | 17.88 | 15 | 98 |
| d25 | 21978 |  |  |  |  |
| *Rural area or Village* | *7068* |  |  |  |  |
| *Small or middle sized town* | *8510* |  |  |  |  |
| *Large Town* | *6396* |  |  |  |  |
| *dk* | *4* |  |  |  |  |
| d63 | 21978 |  |  |  |  |
| *The higher class of society* | *154* |  |  |  |  |
| *The lower middle class of society* | *3456* |  |  |  |  |
| *The middle class of society* | *10942* |  |  |  |  |
| *The upper middle class of society* | *1630* |  |  |  |  |
| *The working class of society* | *5276* |  |  |  |  |
| *Refusal/Other* | *520* |  |  |  |  |

Lastly, country variable is considered. Eurobarometer surveys collected approximately 1000 interviews per country. Only a manipulation is computed: West and East Germany are joined in a single category “Germany”.

To sum up, the final dataset has 21978 respondents (out of 27655). Table 2 shows the number of observations according to country in the original dataset and in the one used in the analysis after data cleaning.

TAB.2: Sample composition

|  |  |  |
| --- | --- | --- |
| **Country** | **Obs in the survey** | **Obs in the analysis** |
| Austria | 1039 | 830 |
| Belgium | 1029 | 970 |
| Bulgaria | 1041 | 626 |
| Croatia | 1016 | 904 |
| Cyprus | 504 | 411 |
| Czech Republic | 1019 | 729 |
| Denmark | 1019 | 839 |
| Estonia | 999 | 520 |
| Finland | 1006 | 807 |
| France | 1029 | 797 |
| Germany | 1539 | 1200 |
| Greece | 1014 | 854 |
| Hungary | 1023 | 900 |
| Ireland | 1067 | 928 |
| Italy | 1022 | 905 |
| Latvia | 1003 | 687 |
| Lithuania | 1001 | 704 |
| Luxembourg | 502 | 399 |
| Malta | 507 | 397 |
| Netherlands | 1022 | 883 |
| Poland | 1013 | 710 |
| Portugal | 1012 | 863 |
| Romania | 1053 | 869 |
| Slovakia | 1076 | 810 |
| Slovenia | 1003 | 874 |
| Spain | 1011 | 820 |
| Sweden | 1034 | 890 |
| United Kingdom | 1052 | 852 |
| **Total** | **27655** | **21978** |

**Analysis**

The following section illustrates the different steps undertaken to obtain a prediction model for pro-environmental action. In particular, the first step consists of Exploratory Data Analysis. Then, the best fitting models tested to predict the final price are presented.

**Exploratory Data Analysis**

**REFERENCES**

Abu-Omar, K., & Rütten, A. (2008). Relation of leisure time, occupational, domestic, and commuting physical activity to health indicators in Europe. *Preventive Medicine*, *47*(3), 319–323. https://doi.org/10.1016/j.ypmed.2008.03.012

Belgiu, M. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 8.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Boutyline, A. (2017). Improving the Measurement of Shared Cultural Schemas with Correlational Class Analysis: Theory and Method. *Sociological Science*, *4*, 353–393. https://doi.org/10.15195/v4.a15

European Commission, Brussels. (2019). *Eurobarometer 91.3 (2019)Eurobarometer 91.3 (2019): Rule of Law, and Climate Change: Rule of Law, and Climate Change* (1.0.0) [Data set]. GESIS Data Archive. https://doi.org/10.4232/1.13372

Fonseca, J. R. S. (2013). Clustering in the field of social sciences: That is your choice. *International Journal of Social Research Methodology*, *16*(5), 403–428. https://doi.org/10.1080/13645579.2012.716973

Goldberg, A. (2011). Mapping Shared Understandings Using Relational Class Analysis: The Case of the Cultural Omnivore Reexamined. *American Journal of Sociology*, *116*(5), 1397–1436. https://doi.org/10.1086/657976

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R* (Vol. 103). Spinger.

Jenhani, I., Amor, N. B., & Elouedi, Z. (2008). Decision trees as possibilistic classifiers. *International Journal of Approximate Reasoning*, *48*(3), 784–807. https://doi.org/10.1016/j.ijar.2007.12.002

Loyen, A. (2016). European Sitting Championship: Prevalence and Correlates of Self-Reported Sitting Time in the 28 European Union Member States. *PLOS ONE*, 17.

Na, S., Xumin, L., & Yong, G. (2010). Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm. *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 63–67. https://doi.org/10.1109/IITSI.2010.74

Peng, C.-Y. J., So, T.-S. H., Stage, F. K., & John, E. P. S. (2002). *THE USE AND INTERPRETATION OF LOGISTIC REGRESSION IN HIGHER EDUCATION JOURNALS: 1988–1999*. 35.

Rossoni, L., Gonçalves, C. P., da Silva, M. P., & Gonçalves, A. F. (2020). *Mapping Organizational Culture Schemes Based on Correlational Class Analysis* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/sf2v4

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer: LOGISTIC REGRESSION: A BRIEF PRIMER. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

1. See table 1 for the list of variables. [↑](#footnote-ref-1)